Отзыв

официального оппонента о диссертации Куликова Сергея Юрьевича на тему: «Автоматическое извлечение мнений: лингвистический аспект», представленной на соискание учёной степени кандидата филологических наук по специальности:10.02.21 – прикладная и математическая лингвистика, М., 2016

При автоматической обработке текстовых массивов существует целый ряд задач, требующих серьёзного рассмотрения. Проблема извлечения знаний отчасти решена, а решение проблемы извлечения мнений сейчас находится в центре внимания специалистов в области прикладной, в частности, компьютерной лингвистики, т.к. понимание того, какое именно мнение выражено в том или ином тексте, позволяет на его основании проводить оценку информации, принимать правильное решение, формировать прогноз.

Тема рецензируемой диссертации Куликова Сергея Юрьевича, несомненно, **актуальна**, так как в современной ситуации информационных войн, реализуемых в СМИ, в том числе и электронных, принципиально важно находить и изучать высказываемое в информационном потоке эксплицитное и имплицитное мнение и оценку происходящих событий, для чего необходима разработка лингвистических принципов автоматического извлечения мнений из текстов различных типов и жанров.

Необходимость подобного исследования определяется отсутствием комплексного описания лингвистического описания извлечения мнений из большого потока текстовой информации, характерного для постоянно расширяющегося информационного поля.

Последовательное изучение теории оценки (Н.Д. Арутюнова, Е.М. Вольф, В.Н. Телия и др.), психолингвистики (И.Н. Горелов, А.А. Леонтьев, Ю.А. Сорокин и др.), теории автоматической обработки текста (Г.Г. Белоногов, Л.Н. Беляева, Ю.Н. Марчук, И.А. Мельчук, А.И. Новиков, Н.В. Лукашевич, R. Schank, Y. Wilks, W. Daelemans и др.), теоретической семантики и синтаксиса (Ю.Д. Апресян, Е.В. Падучева, Л.М. Васильев и др.) и практики автоматического извлечения мнений (А.Н. Соловьев, Т.Е. Загибалов, И.И. Четверкин, J.M. Wiebe, B. Liu, L. Lee, M. Klenner, S. Pulman, M.-T. Taboada и др.) позволило автору работы опираться в своём исследовании на надежный научный базис.

Структура работы традиционна и логична. Она включает введение, три главы, выводы по главам, заключение, список использованной научной литературы и источников, ряд приложений.

Концептуальная цель и соответствующие базовые задачи исследования вполне обоснованно и корректно формулируются автором в следующем виде: изучить особенности существующих систем автоматического извлечения мнений; проанализировать типы оценочной информации, моделируемые в существующих системах; определить классы оценочной лексики, необходимые для повышения качества автоматического извлечения

мнений; уточнить определение понятия «мнение», принятое в практике автоматического извлечения мнений; разработать методы фильтрации априорно нейтрального контента; определить фрагменты этапов автоматического извлечения мнений, зависящие от аспекта задачи; разработать принципы автоматического и автоматизированного создания оценочных ресурсов для русского языка.

Соответственно, **новизна** настоящей работы состоит в том, что автору удалось впервые провести разработку лингвистических принципов автоматического извлечения мнений, призванных повысить качество существующих возможностей идентификации субъективных компонентов контента, учитывая, что действующие системы, построенные на основе статистических моделей без учета собственно лингвистических факторов, недостаточно эффективны. Автором также уточнена классификация оценочной лексики, предложена система из 12 классов. Было доказано, что для оценочной лексики важно ограничение на количество объектов-референтов у оценочных слов, введено понятие однореферентных оценочных слов.

Методы исследования, выбранные автором, соответствуют материалу и общей направленности работы. На основе системного подхода автор в своем исследовании применил методы классификации, корпусного, контекстуального и дискурсивного анализа, метод «черного ящика», статистические методы и метод моделирования, использовались элементы компонентного анализа при исследовании терминологии изучаемой научной области.

Обоснованность научных положений, выводов и рекомендаций, сформулированных в диссертации, их достоверность основана на представительной научно-теоретической базе трудов авторитетных отечественных и зарубежных учёных, анализ которых представлен в первой, второй и частично третьей главах, а также на значительной качественной и количественной выборке материала – тексты сети Интернет (блоги, сообщения информационных агентств и пользовательские отзывы на продукты и события) на русском языке: корпус ruTenTen; корпус русскоязычных СМИ и блогов и ряд других источников (около 150 тыс. словоупотреблений).

Систематизация лингвистической информации, необходимой для задач автоматического извлечения мнений, развитие понятийного аппарата рассматриваемой предметной области определяет **теоретическую значимость** данной диссертационной работы.

Работа значима также в плане детального анализа близких по значению терминов, связанных с извлечением мнений, и мотивированным выбором термина «автоматическое извлечение мнений» в качестве доминантного для этого поля исследований (п.1.10). Структурирование терминологии является принципиально важным показателем теоретического погружения в проблему.

Автор справедливо отмечает, что в последнее десятилетие в

компьютерной лингвистике наметился переход от традиционных узкоспециализированных систем извлечения мнений к гибридным системам, ориентированным на анализ текстов различных стилей и жанров.

Автор также мотивировано определяет понятие «модели анализа» как набора правил сочетаемости оценочных слов и оценочного словаря, которые наиболее точно описывают оценочные отношения конкретной предметной области или набора объектов. По его мнению, ручной выбор модели применяется для узкоспециализированных систем. Для систем общего типа необходим модуль автоматического выбора модели. Автоматический выбор модели анализа осуществляется за счет учета следующих факторов: 1) фактор подъязыка, 2) фактор модели мира, 3) фактор типа объекта. При этом наиболее значимым фактором, по мнению автора, является фактор модели мира.

Не вызывает сомнения утверждение, что для разных задач автоматического извлечения мнений требуется различная организация лингвистического обеспечения и что эти отличия заключаются в наличии или отсутствии дополнительных этапов автоматизированного анализа текстов. Автор выявил, что обязательным компонентом лингвистического обеспечения систем автоматического извлечения мнений является этап фильтрации нерелевантных объектов анализа.

Автор последовательно рассматривает оценочный потенциал вопросительных предложений, придаточных предложений в составе сложного предложения, предложений с анафорой, условных предложений разных типов и пр.

В ходе анализа автором были рассмотрены 3 основных класса систем автоматического извлечения мнений: 1) общего назначения, 2) специализированных и 3) систем идентификации противоправного контента.

Опираясь на результаты анализа материала, автор предлагает общую структуру лингвистического обеспечения системы автоматического извлечения мнений, состоящую из модуля фильтрации контента, модуля лингвистической обработки, модуля автоматического извлечения мнений, ряда подмодулей: фильтрации объектов, определения субъекта, определения объекта, определения оценки. Ключевым компонентом в модуле автоматического извлечения мнений автор справедливо считает подмодуль определения оценки. На вход данного подмодуля поступает список объектов с их синтактико-семантическими характеристиками. На выходе каждому объекту приписываются атрибуты оценки.

Отдельное внимание автор уделяет возможности применения для решения поставленных задач таких современных структур, как онтологии и корпуса текстов с разметкой, отражающей модель мнения, что позволило бы оценивать устойчивость как всей системы в целом, так и ее частей, рассматривая преимущества и недостатки использования онтологий и размеченных корпусов. Автор утверждает, применение онтологий при определении объектов должно быть ограничено узкими предметными областями, список объектов в которых конечен.

Особый интерес вызывают рассуждения автора о категории одушёвленности и ее роли при выделении релевантных субъектов.

Поскольку контент электронных СМИ весьма неоднороден и, можно сказать, нестабилен, полон неадаптированной лексики, автор приходит в выводу, что требуются специализированные базы данных, составление и ведение которых опирается на значительную ручную работу лингвиста. При этом качество автоматического поиска при помощи таких БД будет существенно ниже запросов с использованием усечений (stem*).

Как можно было предполагать, центральным компонентом лингвистического обеспечения системы автоматического извлечения мнения является словарь оценочной лексики. Автор анализирует существующие лингвистические подходы к организации словарей оценочной лексики и их интерпретации, описывает способы представления силы оценки, рассматривает оценочные классы лексики и их представление в словаре, а также в статистических системах. Оценивает возможности учета частеречной и фразеологической информации, описывает правила сочетания оценочных слов, критикует подходы, основанные на машинном обучении.

Серьезное внимание автор уделяет вопросу **практического выхода** диссертационного исследования. На наш взгляд, практическая ценность работы несомненна и заключается в том, что материалы и выводы, полученные в результате данного исследования, могут послужить значимым источником лингвистической информации, а также материалом для повышения качества действующих систем автоматического извлечения мнений за счет совершенствования лингвистического обеспечения. На основе материалов диссертации возможно составление словаря оценочной лексики, формирования баз данных и размеченных корпусов текстов. Кроме того, результаты применимы при разработке комплексной системы автоматического извлечения мнений.

Для подтверждения описанной методики и возможности ее использования на практике был проанализирован корпус отзывов о фильмах (как частный случай отзывов на объекты искусства). На первом этапе был составлен квазисинонимический тезаурус, использована опция автоматического построения тезауруса по поисковому слову, в данном случае «фильм». В результате анализа ошибок методики было выявлено, что фильтрацию низкочастотных слов необходимо приводить после повторной лемматизации с применением морфоанализатора, отличающегося от используемого в системе TreeTagger. После повторной лемматизации автор предлагает произвести пересчет частот. Относительно небольшой объем словаря в дальнейшем предлагается увеличить за счет снижения частотного порога до четырех вхождений в коллокацию со словом из квазисинонимического тезауруса.

В ситуации непрерывно ведущихся информационных и гибридных войн, создания средствами масс-медиа конфликтогенных ситуаций особый интерес вызывает внимание автора диссертации к обработке конфликтогенной лексики, в частности, ксено-, этно- и религиофобонимов.

Вместе с тем, давая в целом высокую положительную оценку научно-теоретическому и прагматико-методологическому вкладу автора диссертации в исследование проблематики автоматического извлечения мнений, компьютерную лексикографию, разработку теории оценки, хотелось бы прояснить следующие дискуссионные моменты:

По сути:

1. В п.1.10 автор исследует существующие отношения в изучаемой терминологии и приходит к выводу, что sentiment analysis и opinion mining являются синонимами. Как, в таком случае, следует относиться к терминам в названиях работ B.Liu "Sentiment Analysis and Opinion Mining" (2012), B.Pang и L.Lee "Opinion Mining and Sentiment Analysis" (2008)?
2. Контент электронных СМИ принципиально неоднороден. Очень часто оценка выражается не только вербально, а со значительным использованием различных лингвокреативных поликодовых элементов, приводящих к высокой степени креолизации текстов, в особенности контента блогов, форумов и пр. При вербальных и орфографических отклонениях предлагается использовать словарь замен, создание которого представляет собой, по нашему мнению, перманентный процесс. Как предполагается обрабатывать креолизованные единицы с оценочной коннотацией?
3. Как автор рассматривает учет сентимент-слов, слов-интенсификаторов и триггеров тональности, таких как «не», «очень» и др.?
4. Как в предлагаемой модели предполагается решать проблемы иронии, сарказма и пр., столь характерные для текстов с оценочным компонентом?

По форме:

1. Излишняя структурная дробность изложения (например, пп. 1.7.1.4 и пр., 2.4.1.3 3 и пр.)
2. Работа несвободна от опечаток (сс. 13, 16, 27 и др.) и неточностей при написании фамилий (Б. Лю и Б. Лиу (с. 83, 84)).
3. Цель Приложений 1-3, где приводятся материалы других авторов, неочевидна, а для научного исследования диссертационного уровня избыточна.

Отмеченные замечания в целом не снижают общей высокой положительной оценки настоящей работы и ее вклада в компьютерную лингвистику и исследования в области искусственного интеллекта, теорию и практику автоматической обработки текстовых массивов, развитие теории структурирования оценочной лексики, в частности, фобонимов и пр., общую и компьютерную лексикографию, теорию и практику автоматического морфологического и семантического анализа.

Рецензируемая диссертация является самостоятельной научно-квалификационной работой. Текст диссертации написан хорошим и ясным русским языком, логично выстроен, обладает внутренним единством.

Работа прошла достаточную апробацию в секторе прикладного языкознания ФГБУН «Институт языкознания Российской академии наук», в докладах автора на международных конференциях и конгрессах, а также в 24 публикациях в сборниках научных трудов, включая три статьи в рецензируемых журналах, рекомендованных ВАК РФ.

Автореферат и публикации автора в достаточной мере отражают содержание настоящей диссертации.

Диссертационное исследование «Автоматическое извлечение мнений: лингвистический аспект» выполнено на высоком научно-теоретическом и методологическом уровне и соответствует требованиям п. 9 Положения о присуждении учёных степеней, а его автор, Куликов Сергей Юрьевич, заслуживает присуждения ему искомой учёной степени кандидата филологических наук по специальности 10.02.21 – прикладная и математическая лингвистика.

Официальный оппонент
доктор филологических наук,
профессор,
профессор кафедры теоретической
и прикладной лингвистики
ГОУ ВО Московский государственный
областной университет — Максименко Ольга Ивановна

16 августа 2016 г.

Ректор МГОУ — П.Н. Хроменков

ГОУ ВО Московской области
Московский государственный областной университет
105005 Москва, ул. Радио, д.10а
kaf-tpl@mgou.ru
(495) 267-89-40